Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process

Luyan Ji[1], Wenfeng Chen[2,4], Tom Loeys[3], & Gilles Pourtois[1]

[1]Department of Experimental-Clinical and Health Psychology, Ghent University, Ghent, Belgium

[2] Department of Psychology, Renmin University of China, Beijing, China

[3]Department of Data Analysis, Ghent University, Ghent, Belgium

[4]Institute of Psychology, Chinese Academy of Sciences, Beijing, China

Correspondence to: Luyan Ji

Email: Luyan.Ji@ugent.be

Abstract

We tested the processing capacity of establishing ensemble representation for multiple facial expressions using the simultaneous-sequential paradigm. Each set consisted of 16 faces conveying a variable amount of happy and angry expressions. Participants judged on a continuous scale the perceived average emotion from each face set (Experiment 1). In the simultaneous condition, the 16 faces were presented concurrently; in the sequential condition, two sets, each containing eight faces, were presented successively. Results showed that judgments varied depending on the number of happy vs. angry faces contained in the sets, and were sensitive at the single trial level to the perceived mean emotion intensity (based on post-experiment ratings), providing evidence of a genuine mean representation rather than the mere use of a single face or enumeration. Experiments 2 and 3 replicated Experiment 1, but implemented a different response format (binary choices) and added masks following each display, respectively. Importantly, in all three experiments, performance was consistently better in the sequential than in the simultaneous condition, revealing a limited-capacity process. A set of control analyses ruled out the use of enumeration, or mere subsampling by the participants to perform the task. Collectively, these results indicate that participants could readily extract mean emotion from multiple faces shown concurrently in a set, but this process is best conceived as being capacity limited.

Keywords: ensemble representation, facial expressions, processing capacity limitations, simultaneous-sequential method, enumeration

Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process

Facial expressions provide important emotional and social signals or cues to guide and optimize communication among human beings. We can infer others' intentions, emotions and attitudes from facial expressions. The processing of emotional faces is sometimes conceived as automatic or capacity-free, requiring minimal attention or even awareness (Tamietto & de Gelder, 2010; Vuilleumier, 2005). However, these initial results have been challenged by other studies showing that processing emotional facial expressions actually requires attention (Pessoa, 2005, 2008; Wolfe & Horowitz, 2004). Research on facial expression perception usually focusses on the processing of faces shown in isolation (i.e., one face at a time; see Calder & Young, 2005; Ekman, 1993; Vuilleumier & Pourtois, 2007). However, in daily life, faces rarely appear in isolation but are surrounded by other faces or objects, such as in an auditorium or at a busy railway station. In addition, unlike single facial expressions, multiple faces are likely to carry mixed social or emotional messages or information, with some of the faces (in the crowd or the audience) that seem happy or pleased, while some other faces may display signs of disapproval or social rejection for example. It remains unexplored and largely unknown whether extracting emotion from multiple faces shown concurrently is subject to processing limitations. The main goal of this study was therefore to explore the possible boundaries of ensemble representation for multiple facial expressions, and eventually assess if this perceptual process is deemed capacity unlimited or not.

Previous studies already demonstrated that human observers can process rapidly and relatively precisely mixed messages or valences (e.g., happy and sad) from multiple faces shown concurrently and in turn extract the average emotion from them (Haberman & Whitney, 2007,

2009; Li et al., 2016). This kind of representation, which combines multiple individual features or items into an emergent quality (i.e., the gist), is referred to as ensemble representation (Alvarez, 2011; Whitney, Haberman, & Sweeny, 2014). Ensemble representations can be formed for a wide range of visual attributes, including both low-level stimuli (e.g., orientation, see Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; size, see Ariely, 2001) and more complex objects (e.g., facial expression and gender, see Haberman & Whitney, 2007).

It is well known that visual perception (and selective attention) is capacity limited (Marois & Ivanoff, 2005). For example, few items can be selected or tracked at once (Scimeca & Franconeri, 2015). However, like low-level features (e.g., size, orientation, contrast), establishing a condensed ensemble representation for higher-order visual information, such as facial expressions, has been proved to be robust and flexible, and is thought to provide an efficient way to overcome or cope with these limited capacity bottlenecks in visual processing (Alvarez, 2011; Cohen, Dennett, & Kanwisher, 2016; Whitney et al., 2014). A main finding supporting this assumption is that the accuracy of ensemble representation remains strikingly high even when individual representations are very poor (impoverished) or even practically lost due to limited attentional resources (Alvarez & Oliva, 2008, 2009; Fischer & Whitney, 2011; Haberman & Whitney, 2009, 2011). The visual system can compensate for noisy local/individual representations by collapsing across those local features to represent the ensemble statistics. For example, when observers were blind to (local) changes in emotional expressions (i.e., they could not precisely localize which face actually changed its facial expression), they could nevertheless accurately report changes in the average emotion of the 16 faces shown in the set (Haberman & Whitney, 2011). Similarly, although participants were unaware of the emotional expression of the central face in the set due to crowding, it nonetheless did impact the perceived average emotion

of the entire set (Fischer & Whitney, 2011). Additional evidence in favor of a capacity unlimited process comes from findings showing that large set sizes yield comparable performance relative to small set sizes (mean emotion, see Haberman & Whitney, 2009; mean size, see Ariely, 2001; Chong & Treisman, 2003), and in some circumstances, performance was even better for the former compared to the latter condition (mean size, see Robitaille & Harris, 2011).

The absence of systematic set size effects is consistent with an unlimited-capacity model whereby processing occurs independently (i.e., without interference or cost) of the number of stimuli shown in the scene or display (i.e., set size). However, the set usually remained relatively homogeneous in spite of varying sizes (i.e., number of individual items). More specifically, similar to earlier psychophysical studies which focused on size processing (Ariely, 2001), Haberman and Whitney (2009) used a uniform distribution of emotional valences composed of only four different facial expressions, no matter which set size was used (i.e., it varied from 4 to 16 faces). In these conditions, observers could presumably sample only a subset of the stimuli, instead of pooling together across all of the individual face stimuli belonging to the set. The use of such a sampling strategy, enabling observers to focus on only one or two sample items regardless of the set size, has been confirmed indirectly by simulation results (Myczek & Simons, 2008; Simons & Myczek, 2008; but cf. Ariely, 2008; Chong, Joo, Emmanouil, & Treisman, 2008; Corbett & Oriet, 2011). To our knowledge, only one study directly examined the subsampling strategy account in the context of facial expression processing, suggesting that sampling four faces (or four faces' worth of information) out of the twelve available in the set could adequately explain participants' fast and precise mean emotion representation, where the contribution of outliers was discounted (Haberman & Whitney, 2010). These strategies could however invalidate the set-size manipulation. When the heterogeneity or variance across items

was maximized, significant set size effects were found in average size perception (Marchant, Simons, & de Fockert, 2013; Utochkin & Tiurina, 2014).

Simultaneous-sequential paradigm

In the present study, we took advantage of the strengths of the simultaneous-sequential paradigm to examine whether extracting the mean emotion from multiple facial expressions obeys the assumptions of an unlimited-capacity process or violates them. The simultaneous-sequential method was initially devised to test the capacity limitations of perceptual processing without the confounds of decision noise induced by set-size manipulations (Eriksen & Spencer, 1969; Shiffrin & Gardner, 1972). In the simultaneous condition, all the individual stimuli were presented at the same time. In the sequential condition, smaller subsets of the stimuli appeared sequentially. Importantly, the duration of each (sub)set was kept constant and thus the amount of time available for processing each stimulus was the same across conditions. Scharff, Palmer and Moore (2011a) introduced a repeated condition, differing only from the simultaneous condition in that the (entire) set appeared twice. Both the unlimited- and the limited-capacity models predict an advantage for the repeated over the simultaneous condition, due to the benefit from viewing the display twice compared with one exposure only. However, under the unlimited-capacity model, divided attention does not affect perception, and all stimuli are processed independently thus there is no interference or competition between them. Since in both the sequential and the simultaneous conditions, each stimulus is displayed for the same duration, the model predicts similar performance between them. By contrast, limited-capacity models predict an advantage for the sequential over the simultaneous condition. This stems from the fact that divided attention over multiple stimuli limits information processing, and only a limited amount

of information can be processed at a given time, thus it can be beneficial to present the stimuli sequentially.

Capitalizing on the extended simultaneous-sequential paradigm, Attarha and colleagues demonstrated that computing multiple ensembles for size or orientation is capacity limited (Attarha, Moore, & Vecera 2014; Attarha & Moore, 2015a; also see Brand, Oriet, & Skyes Tottenham, 2012). However, on the other hand, computing a single ensemble (mean size) of multiple circles was consistent with an unlimited-capacity processing account, and it was the same with average orientation processes (Attarha et al., 2014; Attarha & Moore, 2015a, 2015b), suggesting that summary size and orientation representations appear to be extracted independently for items (e.g., 16 circles or 36 Gabor patches) provided within the single ensembles.

The Present Study

The present study focused on the processing capacity of establishing a single ensemble for multiple facial expressions. High-level ensemble representations (e.g., average facial expressions) have been found to be completely independent from low-level ensemble representations (Haberman, Brady, & Alvarez, 2015). In addition, summary statistics can be established at multiple levels along distinct pathways of the visual hierarchy (Haberman & Whitney, 2012; Hubert-Wallander & Boynton, 2015). Thus, different ensemble representations may engage different processing stages. As reviewed here above, extracting summaries of low-level features (such as size or orientation) embedded in a single set may bypass the limited-capacity processes. However, it remains unclear whether similar effects could be obtained when averaging multiple facial expressions.

Averaging facial expressions requires establishing a high-level ensemble representation, and as such, is a likely candidate for limited-capacity processes. Faces are obviously more complex stimuli or objects, which have multiple dimensions (e.g., configural and discrete features; Rhodes, 2013) compared to attributes like size or orientation, corresponding to low-level or unidimensional features. Using the simultaneous-sequential paradigm, Han & Jung (2016) recently found that even for detecting familiar faces (e.g., the observer's own face or a friend's face), this process was actually capacity limited. However, emotions from facial expressions can be extracted very fast and even under conditions where awareness or attention are massively impoverished, possibly via the (rapid) involvement of subcortical structures and/or multiple parallel routes for visual information processing (e.g., Pessoa & Adolphs, 2010; Tamietto & de Gelder, 2010; Vuilleumier, 2005; Whalen et al., 1998). Hence it is also possible that averaging facial expressions does not require additional steps of converging or integrating component feature populations into a superordinate population code (Attarha & Moore, 2015a), but instead could be established via coarse and fast processing (for example, from the retina to the amygdala, via the superior colliculus and pulvinar, Johnson, 2005; Morris, Öhman, & Dolan, 1999; Vuilleumier & Pourtois, 2007), and as such, may bypass reentrant processing at the cortical level to yield an unlimited processing capacity.

In the present study, we therefore used the extended simultaneous-sequential paradigm (Scharff et al., 2011a) to explore whether establishing ensemble representation for multiple facial expressions depends on limited-capacity processes (i.e., the processing suffers from interference from other stimuli presented simultaneously), or instead, it can be established through unlimited-capacity processes (i.e., the individual stimuli composing the set can be processed independently). To this aim, three experiments were conducted. Participants judged on a

continuous scale the perceived average emotion from each face set conveying a variable amount of happy and angry expressions. The face set consisted of 16 faces (the number of items was chosen to be the same as in Attarha et al., 2014) and was presented for 500 ms (Experiment 1). To examine whether the observed processing limitation was attributed to the response format used in Experiment 1, we ran the same procedure in Experiment 2 but used binary choices (as used in the previous studies, Attarha et al., 2014; Scharff et al., 2011a) instead of the continuous scale. In Experiment 3, masks were presented for 100 ms following each face display for all conditions, in order to prevent any further visual processing of the set after its offset and to rule out potential differences in processing time across conditions. If the performance in the sequential condition (two sets containing each eight faces) had the same accuracy level as in the simultaneous condition (16 faces), we could assume ensemble representations for multiple faces to be capacity unlimited (at least under the experimental conditions under scrutiny). In contrast, if the performance in the sequential condition was better than in the simultaneous condition, we could favor the assumption of limited processing capacities for this grouping mechanism. Importantly, we also performed a set of control analyses (see Methods) to ascertain that participants did not simply use enumeration to comply with task demands, but instead strove to extract the mean emotion from the display containing multiple faces.

General Method

Three different groups of participants were recruited for the three experiments. They all used the same face stimuli and the same two tasks (i.e., average emotion judgment task and face emotion rating task). Before reviewing how they differed from one another, we first present the stimuli and procedure that were common across them.

Participants

All experiments had twenty-four participants from Ghent University (Experiment 1: 18-31 years, 18 females; Experiment 2: 18-26 years, 15 females; Experiment 3: 18-29 years, 21 females). The participants gave written informed consent prior to the start of the experiment and were compensated €10 for their time (1h). They reported to be right-handed and have normal or corrected-to-normal vision. The study protocol was conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee.

Stimuli

Eight male and eight female face identities were selected from NimStim database (Tottenham et al., 2009). Each face identity shows happy and angry expressions. The hair, ears, neck and other external information were cropped. All images were scaled to the same mean luminance and root-mean-square contrast (Bex & Makous, 2002). Each face image subtended a visual angle 4.03° × 4.28°, and was presented against a homogenous black background.

Each set consisted of 16 faces conveying a variable number of happy and angry expressions. These faces were presented in a 4 × 4 invisible grid, spaced horizontally and vertically by 5.29° and 6.46° and centered at the fixation (Figure 1A). The outline of the outer grid (white, 1 pixel) was visible on the fixation screen and face set screen, to help the participants to attend to the part of the visual field where the faces were presented.

Different from previous studies on mean emotion perception where emotion intensity was manipulated continuously using morphing techniques (e.g., Haberman & Whitney, 2007, 2009, 2011), we used full-blown emotional expressions and manipulated the ratio of happy (vs. angry) faces in the set, which was 0.25, 0.375, 0.5, 0.625, or 0.75. Based on these overall ratios,

we determined the configuration of four faces (involving zero, one, two, three or four happy faces) in each quadrant. All the possible combinations were included, except the condition where eight happy or eight angry faces were presented in the two quadrants forming a diagonal plane. Face identities in each set were randomly selected with two constraints: 1) an equal number of male and female faces were presented; 2) the same identity was never repeated in the same set. Depending on the actual ratio of happy vs. angry faces in the sets, a certain number of randomly selected identities were assigned happy expressions, and the rest were assigned angry expressions. The location of each identity in the face set was also randomly determined.

(insert Figure 1 about here)

Apparatus and procedure

Participants sat at 60 cm in front of a 17″ CRT screen, with the position of their eyes roughly aligned to the center of the screen. To minimize head movements, a chinrest was used during the average emotion judgment task. More specifically, participants were asked to judge "what is the average emotion intensity when considering all faces in the set". To this aim, they were encouraged to rely on their first impression and not to think extensively in the average emotion judgment task. When the fixation cross appeared, participants were required to attend to it. After that, participants rated the emotion intensity and arousal of the individual faces. Speed of response was not emphasized and feedback was not given in both tasks. The two tasks were programmed and controlled using the E-Prime Version 2 software (Psychology Software Tools, Inc., 2001). The experiment lasted about 60 minutes.

Average Emotion Judgment Task. The task was derived from the extended simultaneous-sequential paradigm (Scharff et al., 2011a). In the simultaneous condition, the 16

faces were presented concurrently. In the sequential condition, the configuration of face sets was entirely similar to the simultaneous condition mentioned here above, except that they were divided into two subsets, each containing eight faces, and were presented successively. The eight faces were presented along either the positive or negative diagonal (four faces in the upper left and four faces in the lower right; or vice versa). The average emotion of the eight faces in each two diagonal quadrants could be the same or different. Which of the two diagonally opposite positions were presented first in the sequential display was constant for a given participant but counterbalanced across participants (similarly to Attarha et al., 2014, to eliminate uncertainty pertaining to where in the visual field faces were presented). The repeated condition was the same as the simultaneous condition, except that the set of 16 faces was presented twice.

The display type (simultaneous, sequential, repeated) was blocked and the order of them was counterbalanced across participants. The ratio of happy faces was randomized within blocks. Every trial had a unique face set to minimize the visual statistical regularity between trials. After getting acquainted with the average emotion judgment task with 36 practice trials, participants performed three experimental blocks of 120 trials each (24 trials per each ratio). Practice trials were excluded from all subsequent analyses.

A trial began with a fixation cross that was presented in the center of the screen for 500 ms, followed by a face set for 500 ms. In the simultaneous condition, the set with 16 faces was followed by a screen waiting for response (Figure 2A). In the sequential condition, the set with eight faces was followed by a blank ISI of 1000 ms, before the other eight faces was presented along the opposite diagonal, and a response screen (Figure 2B). The repeated condition was the same as the sequential condition except that all 16 faces appeared in both of the two 500- ms

displays (Figure 2C). The next trial automatically began (randomly varying between) 1000 ms-1200 ms after participant responded.

(insert Figure 2 about here)

In order to encourage the participants to process the face set globally and not to focus on a fixed or limited number of faces (or positions) within the set, the procedure also incorporated 24 catch trials (eight trials randomly inserted in each display type), where a red dot (1.36° × 1.45°) unexpectedly replaced one of the faces within the set. Participants were asked to judge the spatial location occupied by the dot and choose from four alternatives: *Upper left*, *Upper right*, *Lower left* or *Lower Right*. They did not need to judge the average emotion when a red dot appeared. The red dot appeared in each of the four quadrants with equal probability in order to foster a broad focus of attention allocated across the whole set. In the sequential and repeated conditions, the appearance of the red dot in the first or the second frame was equiprobable.

Face Emotion Rating Task. Participants evaluated the emotion intensity and arousal of each face previously presented in the average emotion judgment task. One face appeared at a time in the center and had the same size as that in the previous task. Participants used the mouse to click on two different Visual Analogue Scales (VASes). The two anchors of the VAS for emotion intensity were labeled *Extremely positive* and *Extremely negative*. The two anchors of VAS for arousal were labeled *Extremely calm* and *Extremely excited*. The labels on the left and right side were counterbalanced across participants. With this rating task, we could firstly confirm that the happy and angry faces used in this study were perceived as differing in valence, and additionally compute the mean emotion of the sixteen faces in each face set based on this subject-specific emotion intensity ratings obtained for these same faces (see Supplementary

materials), which was used in both the main and control analysis (see General method of data analysis here below). Since a larger variance of items in the set was previously shown to make the averaging task more difficult (Morgan, Chubb, & Solomon, 2008; Solomon, Morgan, & Chubb, 2011), we also computed the variance (standard deviation) of every face set based on these subjective ratings and confirmed that they were similar between the three type conditions (see Supplementary results).

Summary of procedure differences across the three experiments

In Experiment 1, the average emotion judgment task required a response on the VAS. The anchors were labeled *Extremely positive* and *Extremely negative* respectively, which were exactly the same as those used in the face emotion rating task. The displays of the two labels (positive on the left or right) were counterbalanced across participants.

The continuous judgments provide potentially valuable fine-grained information. However, more sophisticated or detailed analyses of the multiple faces might be required in this context when a continuous scale is used, so that establishing a summary statistical representation for them may not happen that early or "automatically". Hence, in Experiment 2, we sought to replicate the procedure of Experiment 1 but used binary choices instead (as used in previous studies, see Attarha et al., 2014; Scharff et al., 2011a) to confirm that the observation of processing capacities for averaging multiple facial expressions (see results of Experiment 1) could not simply be explained by the use of a VAS. Participants were asked to judge whether the average emotion was either positive or negative, by pressing one of the two pre-defined keys on a standard keyboard (either "f" or "j", counterbalanced across participants).

The evidence for a limited-capacity process accounting for the averaging of multiple facial expressions provided in Experiments 1-2 (see the corresponding results sections here below) might be imputed to some uncontrolled differences in terms of stimulus duration between the three different type conditions. Specifically, in the simultaneous condition, the unique display shown was immediately followed (and perhaps this way, partly masked) by the response screen; while in the sequential and the repeated conditions, there was a 1000-ms blank interval following the first display, leading in turn to a potentially longer processing time after its offset in these two conditions. This factor might potentially explain the difference in behavioral performance between the three conditions. To overcome this problem, we replicated Experiment 1, but added masks after each face display in all three conditions (as in Attarha & Moore, 2015a) in Experiment 3. The same mask, a scrambled face image (Figure 1B), was presented for 100 ms immediately following each face display in all three conditions. This mask had the same size as the face stimuli and was presented in the same locations occupied by the faces in the set. We decided to use a VAS as response format in Experiment 3 (like in Experiment 1) because it provided more fine-grained information about the averaging process that we could use directly in some of the control analyses (see here below). Moreover, because results of Experiment 2 (where binary choices were made) were very similar to the results of Experiment 1 (where a VAS was used), we chose to use a VAS in Experiment 3 for comparison purposes. In Experiment 3, we removed catch trials which were used in Experiments 1-2. In these two experiments, catch trials were implemented to encourage a holistic processing of the face set. However since the red dot (used as catch) was salient and thus was quite easy to detect, these catch trials were probably not entirely appropriate to enforce the use of a broad focus of attention. Moreover, these catch trials created additional task demands that we wanted to remove in Experiment 3. Notwithstanding this difference, several control analyses were used in each experiment (see here below) to confirm

indirectly that participants did strive to average the different facial expressions contained in the set, as opposed to focusing on one or two faces only for example.

General method of data analysis

For the average emotion judgment task, the accuracy of catch trials was firstly calculated (Experiments 1-2). The subsequent analyses were based on trials without catch trials.

Data conversion. The actual positions participants clicked on the VASes in the average emotion judgement task were converted to data ranging from 0 to 100 (Experiments 1 & 3). After conversion, the larger the value, the more positive the participants judged the average emotion from the face set; and the smaller this value, the more negative the average emotion from the face set was perceived. We also computed the mean emotion of the sixteen faces in each face set based on the subject-specific emotion intensity ratings (converted in exactly the same way as the average emotion judgment data) obtained for these same faces (see Face Emotion Rating Task here above), and used them in a multilevel statistical model performed as a control analysis (see here below). An absolute difference score was calculated between the average emotion judgment and the computed mean emotion intensity to represent the averaging performance. In Experiment 2 where binary choice was used, we extracted accuracy scores to index the performance. The smaller this difference score and the larger the accuracy score was, the better the averaging performance was.

To directly compare the emotion intensity of happy and angry faces, we subtracted the converted emotion intensity judgment data from 100 for angry faces. Thus, the larger the value, the larger the emotion intensity perceived in the faces by the participants in both cases. The emotion ratings were very similar in the three experiments (Table 1). Paired *t* tests showed that

the perceived emotion intensity of angry faces was stronger than that of happy faces, and angry faces were judged to be more aroused than happy faces, *ps* < .001.

(insert Table 1 about here)

Data trimming. For the average emotion judgment task, trials with RTs faster than 100 ms and exceeding 2.5 SDs above or below the grand mean RT for each participant (overall 2.7%, 5.8%, and 2.6% trials in Experiments 1-3 respectively) were excluded. This standard cutoff was chosen before running data analyses. One participant in Experiment 1 had 19.8% mouse clicks far away from the scale (2.5 SDs above or below the position of the scale), indicating that his/her judgments were often unreliable; hence his/her data were excluded from the analyses. For the remaining participants, 1.1% and 1.7% of trials with mouse clicks far away from the scale were excluded in Experiments 1 & 3, respectively. Because both the unlimited- and the limited-processing capacity models predict an advantage of the repeated condition over the simultaneous condition, another participant in Experiment 1 and one in Experiment 3 who did not show this advantage (i.e., significantly larger absolute difference score in the repeated condition compared with the simultaneous condition) were excluded from further analyses. Note that the same criterion for data trimming was used previously in studies on ensemble representation where a similar paradigm was employed (Attarha et al., 2014; Attarha & Moore, 2015a). The data of the remaining twenty-two, twenty-four, and twenty-three participants were included in the statistical analyses. Noteworthy, adding these participants to the statistical analyses did not change their outcome.

Data analysis. Performance in the average emotion judgment task (the continuous judgment data or the dichotomous accuracy data) was analyzed using repeated-measure

ANOVAs. The two within-subject factors were Type (simultaneous, sequential, repeated) and Ratio (the ratio of happy faces in the set: 0.25, 0.375, 0.5, 0.625, 0.75). Greenhouse-Geisser correction was applied when assumptions of sphericity were violated. A Bonferroni correction was used when multiple comparisons were performed.

Control analyses. A first analysis was conducted to examine the contribution of the four central faces to the average emotion judgments. Two other analyses were carried out to confirm indirectly that participants did not use an extreme or overt subsampling strategy (e.g. focusing on one face only), or mere enumeration to perform the average emotion judgment task.

Average emotion judgment when the mean emotion of the four central faces was neutral. We selected trials for which the four central faces in the set resulted in a mean neutral emotional intensity (33%, 34%, and 34% of all trials in three experiments respectively, two happy and two angry faces in the middle of the set) and analyzed how the average emotion judgments for these specific trials changed with Ratio (the overall ratio of happy faces in the set). This control analysis was carried out to explore the specific contribution of these four central faces to the extraction of the average emotion intensity from the whole display. We reasoned that if participants only focused on them (and merely ignored the 12 other ones shown in the periphery), then their performance would substantially drop (i.e., approach the mid-point on the VAS or the chance level of accuracy) when considering these specific trials only.

Multi-level analyses. Since the average emotion was manipulated by varying the ratio of happy and angry faces in the set, the average emotion was necessarily correlated with the more frequent category present in the set (i.e., when the average emotion was positive, the face set contained more happy faces, and vice versa for an average emotion perceived as negative). In

these standard conditions, it could be argued that the observation of a ratio effect could potentially be explained by the use of an alternative strategy by the participants (compared to the creation of a genuine mean/ensemble representation), namely enumeration or majority-search. In this scenario, participants would merely enumerate the number of exemplars corresponding to a given emotion category or search for the emotion category to which the majority of faces belong, and eventually base their decision on this process, instead of computing the mean emotion from all (or at least most of) faces shown in the set. Thus, to disentangle averaging from enumeration/majority-search strategies, we performed an additional analysis at the single trial level. More specifically, a multilevel model with fixed effects for Type, Ratio (and the interaction between these two factors), and the computed mean emotion intensity (see Face emotion rating task, see also Supplementary materials) as covariate, as well as a random intercept for each subject was fitted for the trial-specific emotion judgments, using SAS PROC MIXED (SAS Institute Inc, 2008; also see Singer, 1998) in Experiments 1& 3 (continuous data) and SAS PROC GLIMMIX (Schabenberger, 2005) in Experiment 2 (binary data, 0 for the positive response, 1 for the negative response), respectively. We reasoned that if one would find in this statistical analysis a significant effect of the (subjective) average emotionality over and above the effect of ratio, this would be consistent with the assumption of the creation of (subject-specific) mean representation in this task, rather than mere enumeration or searching for the majority. We also compared different models with variance of face sets (standard deviation of emotion intensity for 16 faces in each face set, also computed based on the subject-specific emotion intensity ratings) added or with only Ratio or only computed mean intensity involved (Supplementary Table 1).

Response distribution analyses. We also conducted a response distribution analysis to gain a better insight into the meaning of our results. In short, modeling the response distribution

(separately for each condition and experiment) enabled us to assess indirectly whether participants likely used a subsampling strategy, or instead processed the face set as a whole. The details about the rationale of this auxiliary data analysis and the results obtained with it are provided in the supplementary materials section.

Results

Experiment 1

Catch trials. The accuracy of catch trials in the simultaneous (M = .98, SD = .05), sequential (M = .98, SD = .06) and repeated conditions (M = 1.00, SD = 0) was very high.

Average Emotion Judgment. The ANOVA on average emotion judgment revealed a significant main effect of Ratio, F (1.95, 40.84) = 194.69, p < .001, η_p^2 = .90, and an interaction between Type and Ratio, F (8, 168) = 2.31, p = .023, η_p^2 = .10. The effect of Type was not significant, F (2, 42) <1, η_p^2 = .01. When assuming a linear effect of the ratio of happy/angry faces on the participants' judgments, we found evidence for such an effect over the three different conditions, F (1, 21) = 285.82, p < .001, η_p^2 = .93 (Figure 3A). If the face set contained more happy expressions on average, then the participants reliably judged more often the average emotion to be positive (than negative) in this face set. It confirmed that participants' judgments were sensitive to the ratio of happy and angry faces embedded in the set.

Contrast analysis further revealed that the linear effect of ratio was not the same across the three type conditions, F (2, 20) = 5.30, p = .014, η_p^2 = .35. More specifically, the slopes of average emotion judgment in the repeated and the sequential condition were significantly larger than that in the simultaneous condition, ps < .008. On the other hand, there was no significant

difference between these two former conditions, p = .70 (Figure 3A). Participants' judgments were more dispersed and fell on the "wrong" side more frequently in the simultaneous condition (see the response distribution analyses in the supplementary results), so the mean of all judgments in each ratio were closer to the middle point of the scale, leading to a shallower slope. Although a shallower slope usually means a better or a more efficient processing in visual search tasks (Frischen, Eastwood, & Smilek, 2008), here by comparison it reflects a lower sensitivity to the ratio of happy/angry faces in the set and thus a worse performance. We also standardized each average emotion judgment to the mean and standard deviation across all judgments in its specific type condition, to exclude the potential confounds of using scale differently in the three type conditions. The comparisons between types remained the same, however.

For the absolute difference score (Figure 3B), the ANOVA revealed no significant interaction between Type and Ratio, F (8, 168) <1, η_p^2 = .02. There was a significant main effect of Type, F (2, 42) = 11.99, p < .001, η_p^2 = .36. *Post hoc* tests showed that the difference score in the simultaneous condition (M = 18.07, SD = 4.73) was larger than both the repeated (M = 16.52, SD = 3.69) and the sequential condition (M = 15.26, SD = 4.16), p = .026, p = .001, however the difference of the latter two conditions did not reach significance, p = .077. Hence, these results confirmed a limited-capacity model accounting for ensemble representation for multiple emotional facial expressions. There was also a significant main effect of Ratio, F (4, 84) =2.90, p = .027, η_p^2 = .12. The difference scores in the more ambiguous conditions (when the ratio was 0.375, 0.5 and 0.625, M = 17.11, SD = 4.34) were larger than those in the less ambiguous conditions (when the ratio was 0.25 and 0.75, M = 15.86, SD = 3.75), F (1, 21) =5.14, p = .034, η_p^2 = .20.

(insert Figure 3 about here)

Average emotion judgment when the mean emotion of the four central faces was neutral. The average emotion judgment was found to be still reliably influenced by the ratio of angry/happy faces (shown in the periphery). The repeated-measures ANOVA confirmed that there was a significant main effect of Ratio, F (4, 84) = 102.66, $p < .001$, $\eta_p^2 = .83$; although there was no significant main effect of Type, F (2, 42) <1, $\eta_p^2 = .04$, nor interaction effect between Type and Ratio, F (8, 168) = 1.27, $p = .26$, $\eta_p^2 = .06$. These results therefore indicated that the average emotion judgment did not solely depend on the four central faces, but the peripheral faces in the set also contributed to this effect.

Multilevel model analyses. Importantly, we found that at the single trial level, the performance depended not only on the ratio of happy/angry faces contained in the set, but also on the perceived (subject-specific) mean emotion intensity of each face set, as computed based on the post-experiment ratings (Table 2). More specifically, when Ratio and the computed mean intensity were put into the same model, both effects were significant, indicating their reliable contributions to the average emotion judgments. In addition, the Akaike information criterion (AIC, Akaike, 1974), a measure of the relative quality of different statistical models for the given data set, was lower (suggesting a better fit or model) when the computed mean intensity was added together with Ratio, compared with the model including only Ratio as factor. Note that the significant main effect of Ratio and the interaction effect of Ratio and Type in this trial-specific multilevel model were entirely consistent with the outcome of the trial-averaged ANOVA reported here above.

(insert Table 2 about here)

Noteworthy, the significant contribution of the mean emotion intensity (calculated based on all 16 faces contained in the set) to the average emotion judgments did not contradict the limited-capacity account. It did not exclude the likely sampling of a subset of faces in the present case either. Presumably, the mean intensity computed for a smaller number of faces might provide a better fit than the mean based on the 16 individual faces. On the other hand, our results cannot be explained easily by the use of a simple (or extremely economical) strategy that would consist of selecting only one face in the set, or a mere enumeration strategy either. If participants only focused on one face or only relied on the ratio information (enumerate), the specific contribution of the computed mean intensity should be lower and even negligible, which is not what we found nonetheless.[1] Moreover, since the presentation of the set was rather brief (500 ms), (covert) attention was presumably anchored at or close to the fixation point (i.e., the center of the screen), making the central faces the ones that should be selected (and hence contribute to the resulting averaging effect) the most (also see Florey, Clifford, Dakin, & Mareschal, 2016). However, additional analysis controlling for this factor (i.e., the four central faces had a mean neutral emotion) confirmed that the peripheral faces (relative to the central faces in the set) did reliably contribute to the mean emotion intensity extracted from the scene.

Experiment 2

Catch trials. The accuracy of catch trials in the simultaneous (M = .91, SD = .11), sequential (M = .93, SD = .11) and repeated conditions (M = .95, SD = .08) was very high.

[1] The emotion intensity of one face with the strongest expression (either happy or angry) did contribute to the average emotion judgments over and above the ratio. However, the AIC was always lower (suggesting a better fit or model) when the computed mean intensity across 16 faces was added together with Ratio to the model, compared with the model including the intensity of the happiest/angriest face and Ratio.

Average Emotion Judgment. One-sample *t* test showed that the accuracy in each condition was significantly above chance level (*p*s < .001). The repeated-measures ANOVA on accuracy data did not reveal a significant interaction between Type and Ratio, F (6, 138) = 1.39, p = .223, η_p^2 = .06. There was a significant main effect of Type, F (2, 46) = 14.22, p < .001, η_p^2 = .38 (Figure 4B). *Post hoc* tests revealed that the accuracy in the repeated (M = .76, SD = .06) and the sequential conditions (M = .77, SD = .06) were both significantly higher than in the simultaneous condition (M = .69, SD = .07), *ps* < .002. However, there was no significant difference between the repeated and the sequential condition, p > .99. The accuracy results were similar to those of Experiment 1 (see absolute difference scores), and they confirmed a limited-capacity model accounting for ensemble representation for multiple emotional facial expressions.

The main effect of Ratio was also significant, F (1.93, 44.48) = 29.42, p < .001, η_p^2 = .56. Contrast analyses indicated that compared with more ambiguous face sets (when the ratio was 0.375 and 0.625), the accuracy was higher when the mean emotion of faces was presumably less ambiguous (0.25 and 0.75 condition), F (1, 23) = 103.01, p < .001, η_p^2 = .82. In addition, a negativity bias was evidenced since the accuracy in the Ratio0.25 condition (i.e., 75% angry faces) was significantly higher than that in the Ratio0.75 condition (i.e., 75% happy faces), F (1, 23) = 7.75, p = .011, η_p^2 = .25; and additionally, the accuracy in the Ratio0.375 condition (62.5% angry faces) was significantly higher than that in the Ratio0.625 condition (62.5% happy faces), F (1, 23) = 6.36, p = .019, η_p^2 = .22 (Figure 4B).

Average emotion judgment when the mean emotion of the four central faces was neutral. One-sample *t* test showed that the accuracy in each condition for the face sets with a mean neutral emotion in the center was still significantly above chance level (0.5, *p*s < .001).

These results therefore indicated that the average emotion judgment did not solely depend on the four central faces, but the peripheral faces in the set also contributed to this effect.

Multilevel model analyses. Unlike Experiment 1, the subject- specific mean emotion intensity (based on the post-experiment ratings) did not significantly contribute to behavioral performance over and above the effect accounted for by the ratio of happy/angry faces in this experiment (see Table 2). More specifically, when we entered the perceived mean emotion intensity alone (as a unique predictor, without ratio as the competing one), it did significantly predict behavioral performance, $F(1,6935) = 1761.42$, $p < .001$. However, when the effect of ratio was added in the model, the effect of the perceived mean intensity was not significant anymore (Table 2). These results suggest that participants mainly relied on the information about the ratio/number of happy vs. angry faces, rather than perceiving the emotion intensity of those faces in the set. Therefore, we could not exclude the possibility that when only the valence of the average emotion (positive or negative) was used to perform the task, participants mainly used some strategies, like enumerating the number of faces of one emotion category, or searching for the emotion category to which the majority of faces belonged. It is also possible that the emotional expressions of several faces were actually integrated or averaged implicitly (Haberman & Whitney, 2007, 2009), but the simple binary response format used here was not sensitive enough to capture this effect. Nevertheless, the processing capacity was still found to be limited here.

Comparison of Experiment 1 vs. 2. To quantify the effect of response format on the extraction of the average emotion intensity from multiple faces, the continuous judgment data in in Experiment 1 were converted to binary (dichotomous) data to extract accuracy scores and compared them directly to the results of Experiment 2. When the average emotion was positive

(the ratio of happy and angry face was 3:1 or 5:3), if the judgment was larger than 50 (corresponding to the middle of the scale), then we assumed the response to be correct. When the average emotion was negative (the ratio of happy and angry face was 3:5 or 1:3), if the judgment was smaller than 50, then we assumed the response to be correct. When the ratio of happy faces was 0.5, we did not calculate accuracy. The accuracy of average emotion judgments were submitted to a 2 (Response format: continuous vs. binary) × 3 (Type: simultaneous, sequential, repeated) × 4 (Ratio: 25%, 37.5%, 62.5%, 75% happy faces in the set) repeated-measure ANOVA with response format as the between-group factor (Figure 4A). This analysis failed to reveal a significant main effect of Response format, F (1, 44) <1, η_p^2 = .002, or interaction effects including this factor, *ps* > .47. Hence, the two experiments yielded similar results in terms of accuracy, despite changes in the response format used between them, suggesting that establishing a mean representation for facial expressions was best conceived as capacity limited.

We also compared the accuracy of catch trials between the two experiments. Participants performed better in Experiment 1 than Experiment 2, F (1, 44) = 18.25, p < .001, η_p^2 = .29. There were no significant differences between the three display types, however, F (2, 88) = 2.18, p = .12, η_p^2 = .05, nor an interaction between type and response format, F (2,88) <1, η_p^2 = .01.

Experiment 3

Average Emotion Judgment. The ANOVA carried out on average emotion judgments revealed a significant main effect of Ratio, F (1.50, 33.09) = 270.27, p < .001, η_p^2 = .93, an interaction between Type and Ratio, F (5.28, 116.24) = 3.39, p = .006, η_p^2 = .13, but no main effect of Type, F (2, 44) <1, η_p^2 = .04 (Figure 3A).

Similar to Experiment 1, when assuming a linear increase of average emotion judgments as a function of Ratio (ratio of happy/angry faces in the set), we found evidence for such an effect for the three different types, F (1, 22) = 355.48, $p < .001$, $\eta_p^2 = .94$. It confirmed that participants' judgments were overall sensitive to the ratio of happy and angry faces embedded in the set. More specifically, the slopes of average emotion judgments in the repeated and the sequential conditions were significantly larger than that in the simultaneous condition, *ps* < .003. The slopes were similar for the repeated and the sequential conditions, $p = .98$ (Figure 3A). We also standardized each average emotion judgment to the mean and standard deviation across all judgments in its specific type condition, to exclude the potential confounds of using scale differently in the three type conditions. The comparisons between types remained unchanged, and were highly consistent with the results of Experiment 1 (see here below for direct statistical comparison between them).

The ANOVA on the absolute difference scores revealed no significant interaction between Type and Ratio, F (8, 176) = 1.50, $\eta_p^2 = .06$. There was a significant main effect of Type, F (2, 44) = 11.19, $p < .001$, $\eta_p^2 = .34$ (Figure 3B). *Post hoc* tests showed that the difference score in the simultaneous condition (M = 18. 76, SD = 5.29) was larger than both the repeated (M = 17.12, SD = 5.01) and the sequential conditions (M = 16.89, SD = 4.46), $p = .005$, $p = .001$, however the difference of the latter two conditions did not reach significance, $p > .99$, providing additional support for the limited-capacity account to establish an ensemble representation for multiple emotional facial expressions. There was also a significant main effect of Ratio, F (2.56, 56.35) =7.93, $p < .001$, $\eta_p^2 = .27$. The difference score in the Ratio0.25 condition was the smallest, smaller than all the other ratio conditions, *ps* < .019; while there were no significant differences between the other ratio conditions, *ps* > .30.

Average emotion judgment when the mean emotion of the four central faces was neutral. The repeated-measures ANOVA confirmed that there was a significant main effect of Ratio, F (2.21, 44.23) = 96.24, p < .001, η_p^2 = .83; although there was no interaction effect between Type and Ratio, F (5.17, 103.42) <1, η_p^2 = .03. The main effect of Type did not reach significance either, F (2, 40) = 3.07, p = .058, η_p^2 = .13. The average emotion judgment was still reliably influenced by the ratio of angry/happy faces (shown in the periphery). These results therefore excluded the possibility of participants only focusing on the four central faces to carry out the average emotion judgment task.

Multilevel model analyses. Similar to Experiment 1, this analysis showed that the performance depended not only on the ratio of happy/angry faces contained in the set, but also on the perceived (subject-specific) mean emotion intensity of each face set (Table 2). In addition, the model turned out to be better when the computed mean intensity was added together with Ratio, compared with the model including Ratio only. Note that the significant main effect of Ratio and the interaction effect of Ratio and Type in this trial-specific multilevel model were highly consistent with what we found in the standard ANOVA performed on the mean scores obtained for the average emotion judgments (see here above).

Comparison of Experiment 1 vs. 3. The average emotion judgments were submitted to a 2 (Experiment: Exp1 vs. Exp3) × 3 (Type: simultaneous, sequential, repeated) × 5 (Ratio: 25%, 37.5%, 62.5%, 75% happy faces in the set) repeated-measure ANOVA. This analysis failed to reveal a significant main effect of Experiment, F (1, 43) <1, η_p^2 = .004, or interaction effect including this factor, *ps* > .53. Contrast analysis showed that the linear effects of Ratio did not differ between Experiment 1 and Experiment 3 over all three conditions, F (3, 41) < 1, η_p^2 = .02,

and the differences in the slopes between conditions were also the same in these two experiments, F (2, 43) < 1, η_p^2 = .002 (Figure 3A). The mixed ANOVA on the absolute difference scores did not reveal a significant main effect of Experiment, F (1, 43) <1, η_p^2 = .01, or any interaction effect including this factor, *ps* > .07 (Figure 3B). These results indicated that similar results were found in these two experiments and thus that the observation of a capacity-limited process for the averaging of multiple faces was robust and could not be attributed to potential differences in stimulus duration and processing across the three conditions, or the use of catch trials.

General discussion

The present study investigated the processing capacity for establishing an ensemble representation for multiple facial expressions. Consistent with previous studies, the ability to extract the average emotion from multiple emotional faces is deemed very robust and flexible (Haberman & Whitney, 2012). It is valuable because it provides a relatively accurate statistical estimate or summary of a complex visual scene regarding its overall emotional intensity, and probably helps in turn to foster adequate interactions with the social environment. Critically however, based on the use of the stringent extended simultaneous-sequential paradigm, the results of the three experiments reported in this study converge and eventually suggest that this ability is subject to capacity limitations. The performance in the sequential condition (where a smaller set size was at stake for each successive display) was reliably better than in the simultaneous condition where the 16 individual faces were shown at once, which unequivocally supports the limited-capacity model. In fact, performance in the former was equally good as in the repeated condition, suggesting that extracting average emotion likely involves fixed-capacity processing, which can be viewed as an extreme version of the limited-capacity model (Scharff et

al., 2011a). It indicates that only a limited amount of information can probably be processed per time unit, when emotional facial expressions are used.

It has been shown previously that mean emotion from multiple facial expressions could be extracted with limited attention, or awareness (Fischer & Whitney, 2011; Haberman & Whitney, 2011; Ji, Rossi, & Pourtois, under review). In agreement with these earlier results, the current study also unambiguously confirms that averaging of multiple facial expressions can operate efficiently, even though the stimulus presentation was kept short (i.e., 500 ms), and the display contained as many as16 different faces. Notwithstanding this extraordinary perceptual ability, our results also clearly show, based on the use of the simultaneous-sequential paradigm, that extracting mean emotion from a complex set of sixteen faces is capacity limited, however. Attarha et al. (2014, 2015a) previously showed that the accuracy of extracting several means from multiple ensembles in the simultaneous condition was above chance level, but lower than in the sequential and the repeated conditions, consistent with a fixed-capacity model, as evidenced in the current study.

The limited capacity account entails that multiple stimuli could not be processed without interference with each other, and only a limited amount of information can be processed at a given time. In the current study, subsampling strategies (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013; Dakin, 2001; Dakin, Mareschal, & Bex, 2005; Haberman & Whitney, 2010; Myczek & Simons, 2008; Simons & Myczek, 2008) may be at stake during the extraction of the mean emotion from multiple faces, even though its actual nature and modus operandi remain largely unknown. However, it is important to note that even if a complex subsampling strategy was used, it did not necessarily invalidate our main experimental manipulation where we contrasted simultaneous to sequential presentations of the sixteen faces. The comparison between

these two conditions rests on the fact that a smaller number of stimuli were presented in each display in the sequential compared to the simultaneous condition. Importantly, the amount of time available for processing each item was kept constant between these two conditions. It remains somewhat unclear and to be elucidated in future studies whether the use of sampling strategies (implying that a restricted number of stimuli was selected and processed) could account for the capacity limitations, or the other way around, such subsampling strategy derives from the fact that capacity limitations prevail when the average emotion has to be computed from sixteen different and briefly presented faces.

In the sequential condition, two averages had to be computed and later integrated with one another, creating an extra averaging component that was not present in the simultaneous condition, which might also contribute to the differences found in the averaging performance between these two conditions. However, the better performance in the sequential compared to the simultaneous condition found in the current study could not be attributed merely to the benefits created by averaging two separate displays. Instead, we contend that capacity limitations in extracting the mean emotion most likely accounted for this difference between them. If we assume that all items could be processed independently (i.e. an unlimited capacity process), then observers would average the same amount of information in the simultaneous and the sequential conditions, making the averaging process very similar for them and hence leading to a statistically undistinguishable behavioral performance between them. Alternatively, if we reckon that capacity limitations restrict information processing, when the putative benefit of integrating two displays is even removed, such as in visual search tasks where no averaging is required, the performance is still better in the sequential compared to the simultaneous condition (Han & Jung, 2016; Scharff, et al., 2011a, 2011b), as we have found here in this study where averaging was

required. Accordingly, integrating two separate averages in the sequential condition probably granted an additional advantage to this condition over the simultaneous condition, pending the averaging process was limited however.

In spite of focusing on a limited amount of information in the set (or having interference between multiple individual items), a rather precise mean estimate could be computed, as our new results show. One reason accounting for this paradox might be that the set was statistically regular (Alvarez, 2011). The same face identities were repeated across trials and conditions, although with changing locations and emotional expressions each time, hence inevitably creating some redundant information and statistical regularity. The co-activation model suggests that neural signals from multiple redundant stimuli are summed up (Miller, 1982), and redundant faces have been found to facilitate perception by enhancing the robustness of representation of each face (Won & Jiang, 2013). Although there are still controversies whether redundant information is compressed or not (Baijal, Nakatani, van Leeuwen, & Srinivasan, 2013; Brady & Alvarez, 2011), the selected subset, though possibly biased, provides a reliable estimate or proxy of the whole set to some extent.

In contrast to previous studies which primarily used morphed faces of one single person/identity (created by interpolating/blending between two different emotional expressions, see Haberman & Whitney, 2007, 2009, 2012), here we employed face images of different identities conveying natural expressions (similar to Simmons, Stein, Matthews, Feinstein, & Paulus, 2006; Yang, Yoon, Chong, & Oh., 2013). The advantage of using these different facial expressions from different people is that they have higher ecological validity. After all, it is rather odd to judge one person's various emotional expressions at the same time, but more common and reasonable to judge the overall emotion of a crowd based on multiple individuals' emotional

expressions. A downside of this approach however is that we did not have *objective* measures of emotional intensity for each face shown in the sets (e.g., morphing values, Haberman & Whitney, 2007). Nevertheless, to overcome this limitation, we collected post-experiment ratings of these faces that in turn provided *subjective* estimates of their emotion intensity. After all, emotion intensity from facial expressions can hardly be captured by an arbitrary numerical value being constant for all subjects, but most likely is dependent upon the specific experiment-dependent and viewer-specific conditions.

Another caveat of our approach is the lack of independence between the ratio manipulation on the one hand and the dependent variable on the other hand. As a matter of fact, the average emotion in the set was inevitably correlated with the more frequent emotion category included in this set. In this context, strategies of enumerating or actively searching for the more frequently occurring category might be used, instead of extracting the mean emotional information based on an averaging process. To rule out this possibility more formally, we conducted multilevel model analyses (at the single trial level) with the computed subject-specific mean intensity included as predictor (besides the ratio effect) for all three experiments. The results of Experiments 1 & 3 for these multilevel model analyses unequivocally confirmed that besides ratio per se, the mean emotion intensity did reliably account for behavioral performance during the task. In fact, the model provided each time the best fit when these two factors were included together in the statistical model, suggesting thereby that the averaging performance could hardly be explained by the use of mere enumeration of the emotional faces. When participants were required to judge both the valence and the intensity of the mean emotion in Experiments 1 & 3 (where a continuous response format was used), results showed that enumeration or relying merely on the ratio information could not satisfactorily account for them.

Interestingly, although the strongest (either happy or angry) face in the set reliably predicted the averaging performance over and above ratio (see Footnote), this statistical model was not fitted as good as the one including the mean intensity of all faces in the set as factor. These results suggest that at least two or more were sampled and integrated in the averaging, which fulfills the criterion of ensemble representation according to recent models (Whitney & Leib, 2018). Therefore, as the most parsimonious interpretation, it appears that participants did strive to integrate multiple facial expressions and form an ensemble representation in our study, as they were explicitly asked to do. Noteworthy, the resulting mean emotions were necessarily approximations, instead of an exact sum divided by the number of faces in the set, but as our results show, it seems that they represented the (sampled) set as a whole rather precisely (Chong et al., 2008).

Nevertheless, in Experiment 2 where binary choices were used, the computed subject-specific mean emotion intensity did not significantly contribute to the performance over and above the effect of ratio of happy/angry faces, and thus we could not formally claim that mean emotion intensity was extracted in this experiment. On one hand, given the task (i.e., simple discrimination), the processing for multiple facial expressions was presumably more coarse or superficial in Experiment 2 than Experiments 1 & 3. In Experiment 2, participants did not really need to perceive the emotion intensity of these faces to carry out the task, which was different from what was required in Experiments 1 & 3 where a continuous scale was implemented and thereby what we found in these two experiments using this specific data analysis. The lower accuracy for catch trials (although the performance was still very good) in Experiment 2 compared with Experiment 1 also suggested indirectly that participants might have used a more local processing strategy (e.g., mostly relying on ratio information, or subsampling) in the former compared to the latter. Possibly, when binary choices were used (Experiment 2), a coarse

processing of the sets was sufficient. On the other hand, the higher accuracy score in the sequential than the simultaneous condition found in Experiment 2 also supported a limited-capacity account. The specific contribution of enumeration or majority-search in processing multiple facial expressions remains to be fully elucidated and should be carefully controlled in future studies, including using specific control analyses as carried out here. For example, this could be achieved by manipulating both the ratio of different kinds of expressions and the mean intensity of faces in the set (e.g., the average emotion is more positive even when there are more negative faces in the set), like previously done in a study focusing on gender processing (Nagy, Zimmer, Greenlee, & Kovács, 2012).

An important unanswered question in our study relates to what factor(s) eventually determines exactly these limitations in ensemble representation for facial expressions. Previously, the processing of emotional facial expressions (when used in isolation), especially threatening ones such as angry or fearful faces, was reported to occur rapidly and take place under conditions characterized by impoverished awareness or limited attention, presumably through a fast subcortical pathway (Tamietto & de Gelder, 2010; Vuilleumier, 2005; Whalen et al., 1998). Traditional visual search tasks also found that searching for a single negative (angry) face from multiple (face) distractors was also very efficient, and probably consistent with the use of a parallel/automatic attention process (Frischen et al., 2008). However, these results do not necessarily generalize to conditions where multiple faces are used, and all need to be processed concurrently, such as required here. It is possible that discriminating a group of multiple faces all belonging to one emotion category (“anger”) from another group of faces with another emotion content (“happiness”) is actually capacity limited. Enumerating a larger number of items (e.g., over 4 items), which might occur in our case if participants possibly relied on the enumeration

strategies (Experiment 2), engages a limited-capacity (Trick & Pylyshyn, 1994). The potential strategy of majority search (whether there are more happy or more angry faces, Fan, Guise, Liu, & Wang, 2008) is found to involve voluntary cognitive control as well. Furthermore, we cannot exclude the third possibility that the averaging process itself, when multiple emotional facial expressions are used in a single set (unlike lower level features such as size or orientation), is inherently capacity-limited, probably because face stimuli are more complex and are in essence multi-dimensional objects (e.g., first-order and second-order features; identity and expressions; valence and intensity; viewpoint). These possibilities mentioned here above might not be mutually exclusive. Accordingly, future studies using different types or combinations of emotions (for example, neutral and angry/fearful, or angry faces with different intensities), as well as using possibly simpler face stimuli (e.g., schematic faces) are needed to corroborate first the assumption that establishing an ensemble representation for multiple facial expressions is in essence capacity limited, and next to examine the specific perceptual or attentional processes responsible for these capacity limitations, informing in turn about the boundaries of this process. In addition, whether or not the processing capacity might differ between high-level versus low-level ensemble representation is also an interesting avenue for future research, using preferably within-subject experimental designs enabling a direct statistical comparisons of the averaging process between them.

In sum, the results from three separate experiments gathered this study concur and support the idea that although human observers could rapidly extract the affective gist of multiple faces with different identities and variable emotional expressions shown concurrently and briefly, ensemble coding of multiple facial expressions is characterized by capacity limitations. Importantly, several control analyses also showed that the use of only one face or mere

enumeration by the participants were unlikely to explain these results. There are clear boundaries regarding its dependence upon built-in attention resources or processes, confirming the assumption that sampling complex visual scenes with the aim to extract their affective gist likely requires the involvement of additional attention processes, and thus specific feedback or re-entrant processing in the visual cortex (Di Lollo, Enns, & Rensink, 2000; Lamme & Roelfsema, 2000; Pessoa, 2008).

Acknowledgement

This work is supported by a China Scholarship Council (CSC) grant ([2014]3026) and a cofounding grant from Ghent University, both awarded to LJ, and a grant from the National Natural Science Foundation of China (31371031) awarded to WC.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, *83*, 25-39.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122-131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392-398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*(18), 7345-7350.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157-162.

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics*, *70*(7), 1325-1326.

Attarha, M., & Moore, C. M. (2015a). The capacity limitations of orientation summary statistics. *Attention, Perception, & Psychophysics*, 1-16.

Attarha, M., & Moore, C. M. (2015b). The perceptual processing capacity of summary statistics between and within feature dimensions. *Journal of Vision, 15*(4), 9, doi:10.1167/15.4.9.

Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary statistics of size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensembles.

Journal of Experimental Psychology: Human Perception and Performance, 40(4), 1440-1449.

Baijal, S., Nakatani, C., van Leeuwen, C., & Srinivasan, N. (2013). Processing statistics: An examination of focused and distributed attention using event related potentials. *Vision Research, 85,* 20-25.

Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of the Optical Society of America A, 19*(6), 1096-1106.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science, 22,* 384-392.

Brand, J., Oriet, C., & Sykes Tottenham, L. (2012). Size and emotion averaging: Costs of dividing attention after all. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(1), 63.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*(8), 641-651.

Chong, S. C., Joo, S. J., Emmmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70*(7), 1327-1334.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences, 20*(5), 324–335.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica, 138*(2), 289-301.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A, 18*(5), 1016-1026.

Dakin, S. C., Mareschal, I., & Bex, P. J. (2005). Local and global limitations on direction integration assessed using equivalent noise analysis. *Vision Research, 45*(24), 3027-3049.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, *129*(4), 481.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384.

Eriksen, C. W., & Spencer, T. (1969). Rate of information processing in visual perception: Some results and methodological considerations. *Journal of Experimental Psychology, 79*(2p2), 1.

Fan, J., Guise, K. G., Liu, X., & Wang, H. (2008). Searching for the majority: Algorithms of voluntary control. *PLoS One*, *3*(10), e3522.

Florey, J., Clifford, C. W., Dakin, S., & Mareschal, I. (2016). Spatial limitations in averaging social cues. *Scientific Reports, 6*.

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, *106*(3), 1389-1398.

Frischen, A., Eastwood, J. D., & Smilek, D. (2008). Visual search for faces with emotional expressions. *Psychological Bulletin*, *134*(5), 662.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751-R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 718.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825-1838.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*(5), 855-859.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From perception to consciousness: Searching with Anne Treisman*, 339-349.

Han, S. W., & Jung, W. H. (2016). Your own face is no more precious than others': Evidence from the simultaneous–sequential paradigm. *Psychonomic Bulletin & Review*, *23*(1), 187-192.

Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, *15*(4), 5, doi: 10.1167/15.4.5.

Johnson, M. H. (2005). Subcortical face processing. *Nature Reviews Neuroscience*, *6*(10), 766-774.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571-579.

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, *7*.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142*(2), 245-250.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences, 9*(6), 296–305.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247-279.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of Vision, 8*(11), 9, doi: 10.1167/8.11.9.

Morris, J. S., Öhman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences, 96*(4), 1680-1685.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772-788.

Nagy, K., Zimmer, M., Greenlee, M. W., & Kovács, G. (2012). Neural correlates of after-effects caused by adaptation to multiple face displays. *Experimental Brain Research, 220*(3-4), 261-275.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739-744.

Pessoa, L. (2005). To what extent are emotional visual stimuli processed without attention and awareness?. *Current Opinion in Neurobiology, 15*(2), 188-196.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148-158.

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience, 11*(11), 773-783.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12), 18, doi: 10.1167/11.12.18.

Rhodes, G. (2013). Looking at faces: first-order and second-order features as determinants of facial appearance. *Perception, 42*(11), 1179-99.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schabenberger, O. (2005). Introducing the GLIMMIX procedure for generalized linear mixed models. *SUGI 30 Proceedings*, 196-30.

Scharff, A., Palmer, J., & Moore, C. M. (2011a). Extending the simultaneous-sequential paradigm to measure perceptual capacity for features and words. *Journal of Experimental Psychology: Human Perception and Performance, 37*(3), 813.

Scharff, A., Palmer, J., & Moore, C. M. (2011b). Evidence of fixed capacity in visual object categorization. *Psychonomic Bulletin & Review, 18*(4), 713–721.

Scimeca, J. M., & Franconeri, S. L. (2015). Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 109–118.

Shiffrin, R. M., & Gardner, G. T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology, 93*(1), 72.

Simmons, A., Stein, M. B., Matthews, S. C., Feinstein, J. S., & Paulus, M. P. (2006). Affective ambiguity for a group recruits ventromedial prefrontal cortex. *Neuroimage, 29*(2), 655-661.

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics*, *70*(7), 1335-1336.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*(4), 323-355.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, *11*(12), 13, doi: 10.1167/11.12.13.

Tamietto, M., & de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience*, *11*(10), 697-709.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... & Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242-249.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*(1), 80.

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and de Fockert. *Acta Psychologica*, *146*, 7-18.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585-594.

Vuilleumier, P., & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia*, *45*(1), 174-194.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, *18*(1), 411-418.

Whitney, D., Haberman, J., & Sweeny, T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. In J.S. Werner and L.M. Chalupa (Eds.), *The New Visual Neurosciences* (pp.695-710). MIT Press.

Whitney, D., & Leib, A. Y. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 12.1-12.25.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it?. *Nature Reviews Neuroscience*, *5*(6), 495-501.

Won, B. Y., & Jiang, Y. V. (2013). Redundancy effects in the processing of emotional faces. *Vision Research*, *78*, 6-13.

Yang, J. W., Yoon, K. L., Chong, S. C., & Oh, K. J. (2013). Accurate but pathological: Social anxiety and ensemble coding of emotion. *Cognitive Therapy and Research*, *37*(3), 572-578.

Table 1

Summary of Emotion Ratings in Experiments 1-3

	Angry_Intensity	Happy_Intensity	Angry_Arousal	Happy_Arousal
Exp1	83.29(5.87)	75.16(5.06)	53.40(4.53)	32.74(5.69)
Exp2	82.15(6.29)	70.25(4.86)	61.85(5.39)	41.22(6.72)
Exp3	81.56(7.79)	70.24(5.57)	61.08(7.45)	36.43(10.28)

Note. The intensity and arousal rating (means and standard deviations) for the faces in Experiments 1-3. The perceived intensity of angry faces was stronger than that of happy faces, and angry faces were judged to be more aroused than happy faces, in all three experiments.

Table 2

Summary of Multilevel Models to Predict Average Emotion Judgments in Experiments 1-3

	Type	Ratio	Type * Ratio	Mean Intensity
Exp1	$F(2, 7586) < 1$	$F(4, 7586) = 20.52$, $p < .001$	$F(8, 7586) = 2.51$, $p = .010$	$F(1, 7586) = 37.50$, $p < .001$
Exp2	$F(2, 8094) = 2.62$, $p = .073$	$F(4, 8094) = 41.08$, $p < .001$	$F(8, 8094) = 6.48$, $p = .001$	$F(1, 8094) = 1.99$, $p = .158$
Exp3	$F(2, 7886) = 1.48$, $p = .228$	$F(4, 7886) = 50.35$, $p < .001$	$F(8, 7886) = 3.47$, $p < .001$	$F(1, 7886) = 4.98$, $p = .026$

Note. The predictors include Type (simultaneous, sequential and repeated display type), Ratio (25%, 37.5%, 50%, 62.5% and 75% happy faces in the set), the interaction between Type and Ratio, and Mean intensity (computed for each face display and for each participant based on their post-experiment emotional ratings). Experiments 1 and 3 confirmed that the continuous average emotion judgments depended not only on the ratio, but also on the perceived mean emotion intensity of each face set. However, Experiment 2 showed that the perceived mean intensity did not significantly contribute to the binary average emotion judgments over and above the effect accounted for by the ratio of happy/angry faces.

Figure Legends

Figure 1. Face set used in Experiments 1 - 3. The average emotion of the set in this example, including 25% angry faces and 75% happy faces, is positive.

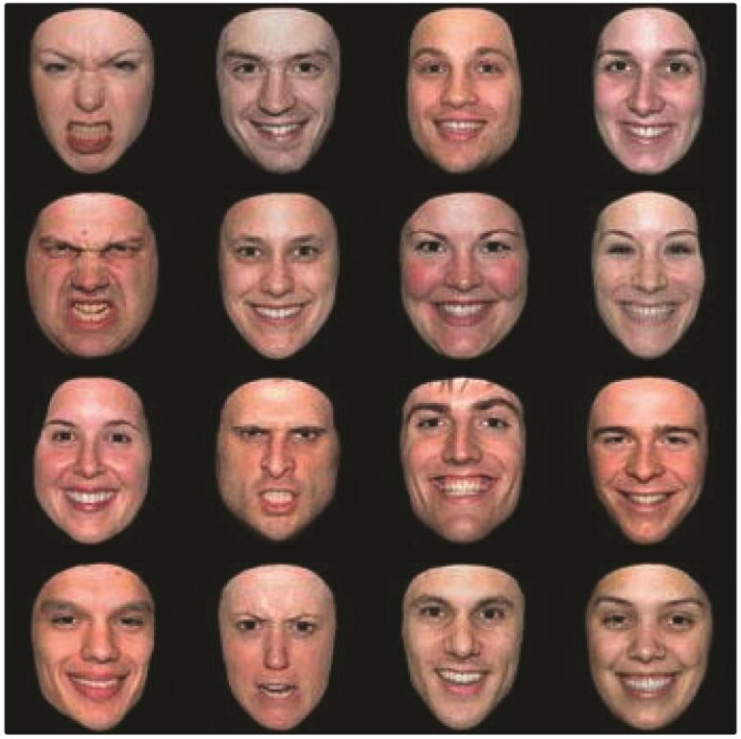
Figure 2. Average emotion judgment task. Trial events for the (A) simultaneous, (B) sequential, and (C) repeated conditions in Experiment 1. Instead of schematic faces (used for illustration purpose), real faces (photographs) were presented against a black background. Participants judged the perceived average emotion intensity from each face set on a visual analogue scale, ranging from extremely negative to extremely positive (these two anchors were counterbalanced across participants). The next trial began (randomly varying between) 1000 ms – 1200 ms after participants responded.

Figure 3. (A) Average emotion judgment (means) and (B) absolute difference scores (means) between the average emotion judgment and the computed mean emotion intensity shown separately for the five different ratios and the three different display types used in Experiment 1 (Upper) and Experiment 3 (Lower). The larger the judgment, the more positive participants perceived the face set; the smaller the judgment, the more negative participants judged it. The column graphs show the slopes of the average emotion judgment (means) in the three conditions. The larger the absolute difference scores, the worse the performance. ‘SI’ stands for simultaneous, ‘SE’ for sequential, and ‘RE’ for repeated condition. The error bar represents one standard error of mean.

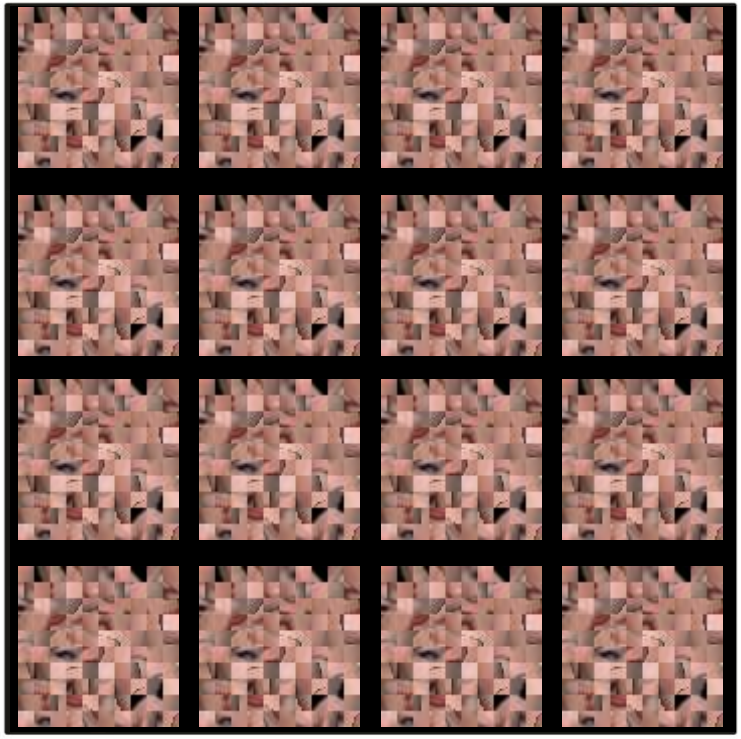
Figure 4. Accuracy of average emotion judgment (means) shown separately for the five different ratios and the three different display types used in Experiment 1 and Experiment 2. ‘RE’

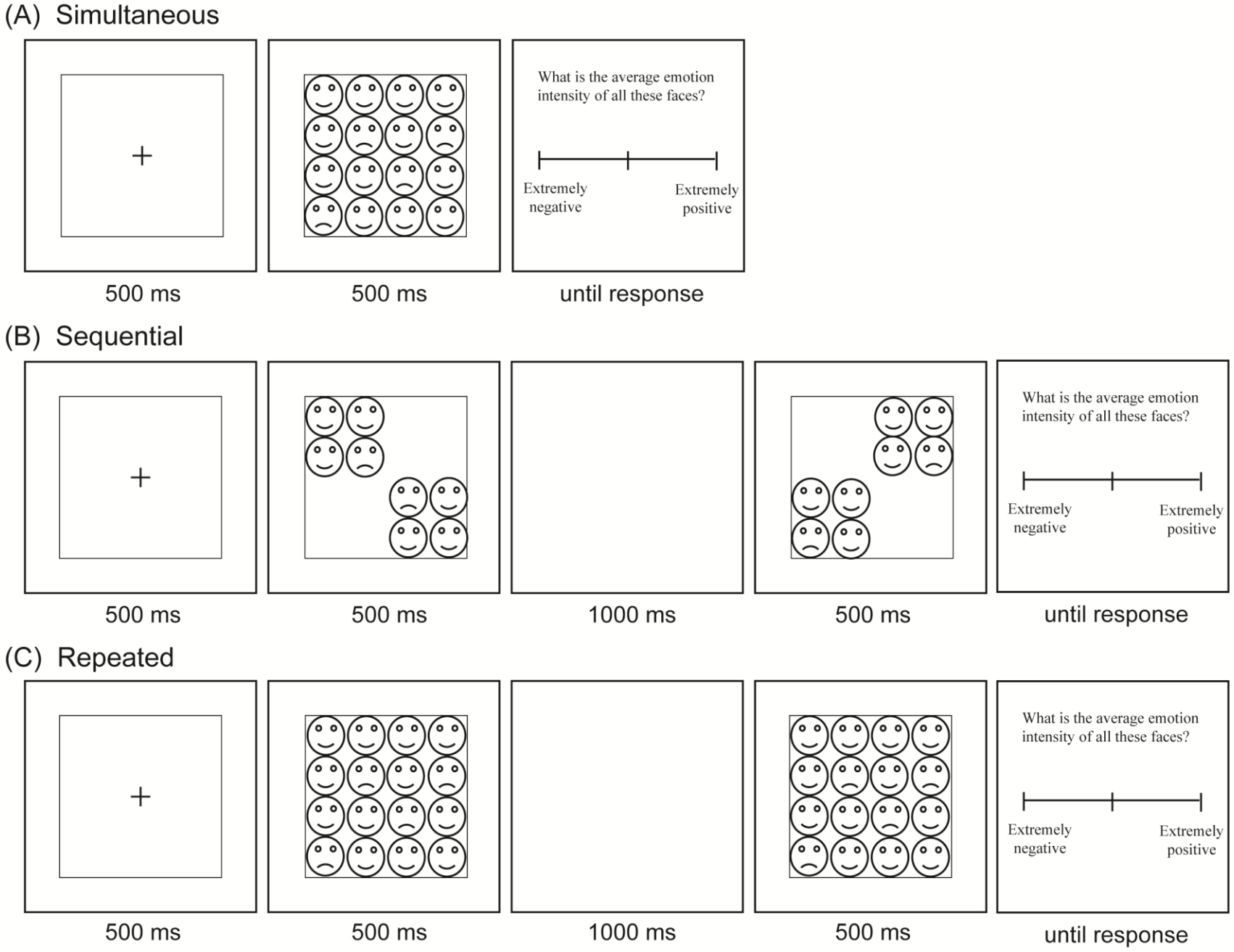
stands for repeated, ‘SE’ for sequential and ‘SI’ for simultaneous condition. The error bar represents one standard error of mean.

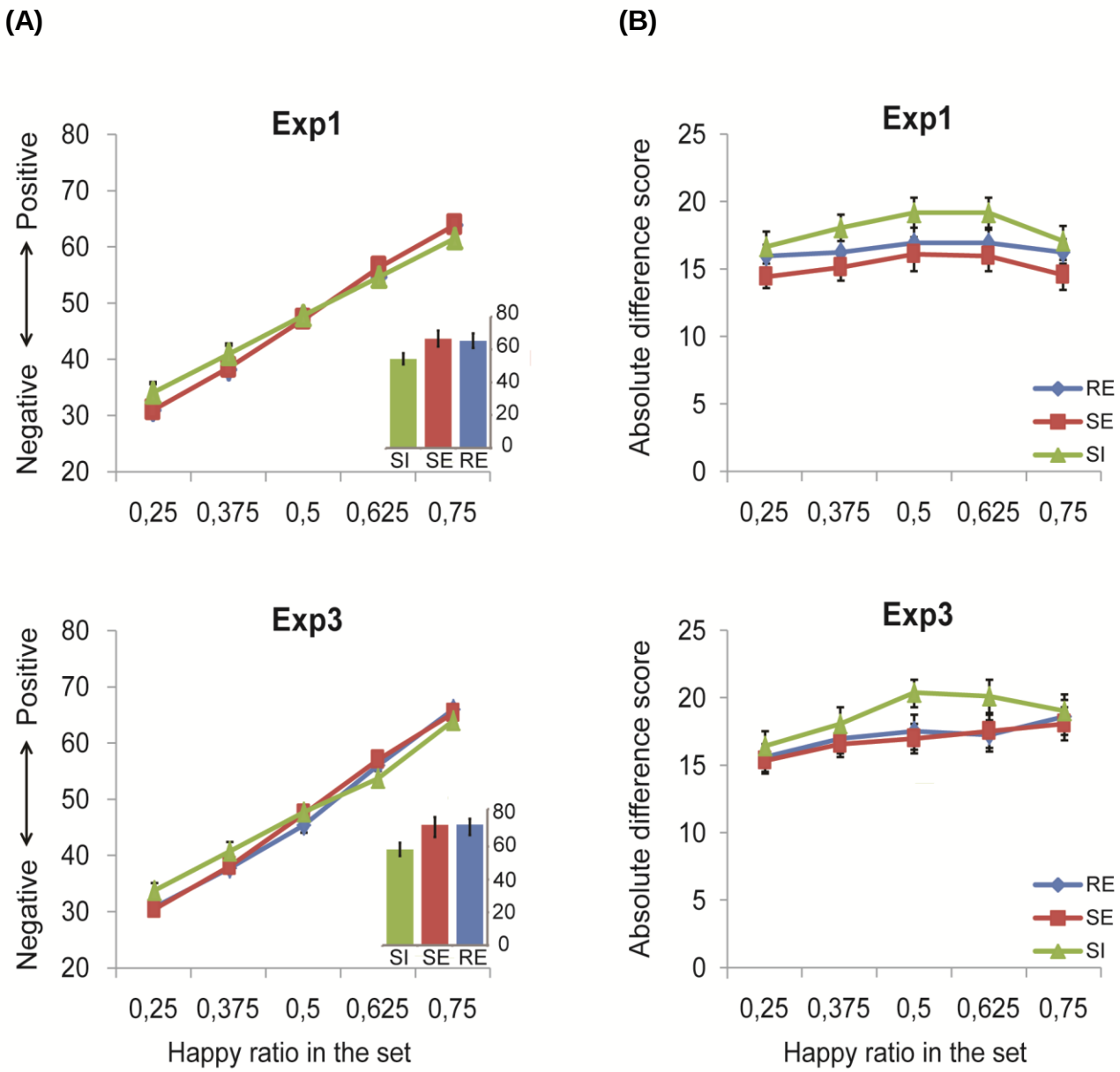
(A)



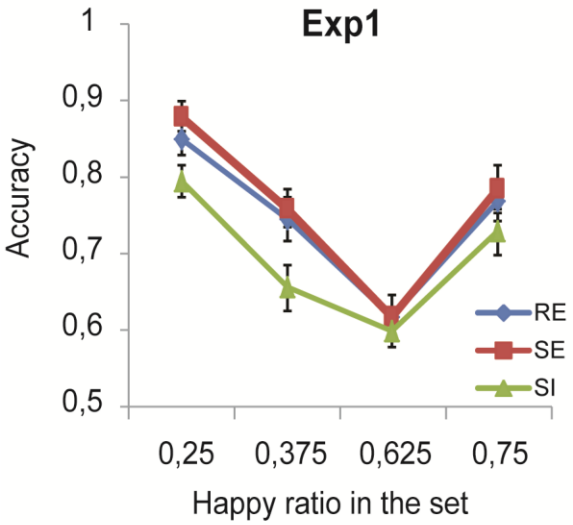
(B)



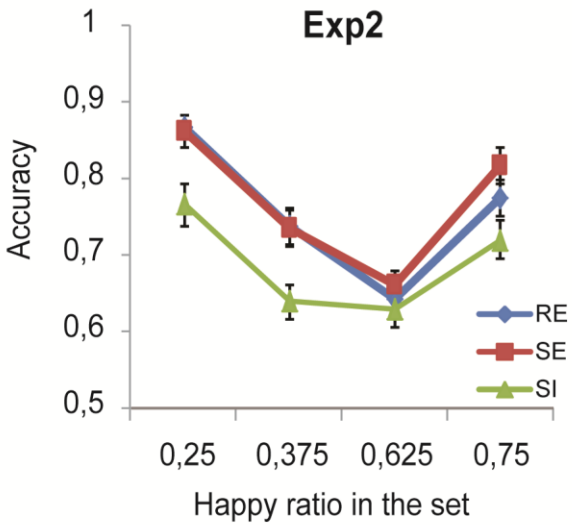

(A) Simultaneous
What is the average emotion intensity of all these faces?
Extremely negative
Extremely positive
500 ms
500 ms
until response
(B) Sequential
What is the average emotion intensity of all these faces?
Extremely negative
Extremely positive
500 ms
500 ms
1000 ms
500 ms
until response
(C) Repeated
What is the average emotion intensity of all these faces?
Extremely negative
Extremely positive
500 ms
500 ms
1000 ms
500 ms
until response


(A)
(B)
Exp1
Exp3
Negative ⟷ Positive
Absolute difference score
80
70
60
50
40
30
20
25
15
10
5
0
0,25 0,375 0,5 0,625 0,75
Happy ratio in the set
SI SE RE
RE
SE
SI

(A)


Exp1
1
0,9
0,8
0,7
0,6
0,5
Accuracy
RE
SE
SI
0,25
0,375
0,625
0,75
Happy ratio in the set

(B)


Exp2
1
0,9
0,8
0,7
0,6
0,5
Accuracy
RE
SE
SI
0,25
0,375
0,625
0,75
Happy ratio in the set

Supplementary Materials

Supplementary Results

Mean emotion intensity of face sets. We computed the arithmetic mean emotion of the sixteen faces in each set based on the subject-specific emotion intensity ratings obtained for these same faces. The larger the value, the more positive the computed mean emotion intensity was, while conversely, the smaller this value, the more negative the computed mean intensity was. As can be seen from the supplementary Figure 1A-C, the mean emotion intensity calculated based on individual intensity ratings for each face had large variance, but was still clearly sensitive to the ratio of happy versus angry faces contained in the set. Linear regression analyses confirmed this observation. For each experiment separately, we calculated a simple linear regression to predict the computed mean intensity based on Ratio, and significant effects were found for all three experiments, Exp1: F (1, 7625) = 32.01, p <.001, adjusted R^2 = .81; Exp2: F (1, 8136) = 29.58, p < .001, adjsuted R^2 = .78; Exp3: F (1, 7922) = 31.30, p <.001, adjusted R^2 = .80.

(insert Supplementary Figure 1 about here)

Variance of face sets. When using two opposite expressions (happy and angry, as here in our study), the different face sets were highly variable compared with earlier studies using morphed faces (e.g., ±3 and ±9 emotional units around a randomly selected set mean; see Haberman & Whitney, 2007, 2009). An increased variance across individual sizes or orientations in the sets was previously shown to make the averaging task more difficult (Morgan, Chubb, & Solomon, 2008; Solomon, Morgan, & Chubb, 2011). In the current study, we also computed the variance (standard deviation) of every face set for each participant based on their subjective emotional intensity ratings, and conducted a repeated-measure ANOVA with Type (Simultaneous, Sequential, Repeated) and Ratio (25%, 37.5%, 50%, 62.5%, 75% happy faces in the set) as two within-subject variables, and with Experiment (Experiments 1, 2, & 3) as the between-subject

variable. There was no interaction between any of the two or three factors, *ps* > .63. The main effect of Ratio was significant, F (1.86, 122.50) = 481.74, $p < .001$, $\eta_p^2 = .88$. *Post hoc* tests revealed that the more ambiguous the face sets were, the larger the variance became. When there were 50% happy faces (50% angry faces) in the set, the variance of the set was the highest (*p*s < .001), while when there were 25% or 75% happy faces in the set, the variance of the set was the lowest (*p*s < .001). The variance of the set with 37.5% or 62.5% happy faces lied between these two conditions (*p*s < .001). In addition, the symmetrical conditions did not differ from each other (Ratio0.25 vs. Ratio0.75, Ratio0.375 vs. Ratio0.625), *p*s > .89. The main effect of Type was also significant, F (2, 132) = 3.12, $p = .048$, $\eta_p^2 = .05$, however *post hoc* tests did not find significant differences between any of the three type conditions (SI: M = 28.19, SD = 5.92; SE: M = 28.13, SD = 5.93; RE: M = 28.20, SD = 5.96), *p*s > .11.

The variance of face sets (albeit being similar between the three main conditions) was higher when the face sets were more ambiguous in terms of mean emotion displayed (i.e., the ratio of happy/angry faces in the set became less asymmetrical). However, when we used this metric as predictor in our multilevel analyses, it did not contribute significantly to explain behavioral performance (unlike the perceived and subject-specific mean emotion). More specifically, when the ratio was included as predictor in the model, the effect of variance was never significant (see Supplementary Table 1). Accordingly, we can conclude that item variability did not contribute as much as mean emotion intensity in guiding performance when there were 16 faces shown in the set.

Response distribution analyses. While the previous control analyses already ruled out somehow the systematic use of a simple subsampling strategy, we also wanted to provide additional and more potent evidence for this conclusion. Since there were different ratios of

happy vs. angry faces across conditions in the set, obviously, the probability of participants seeing a single happy or a single angry face also changed with those ratios. If a participant would exclusively process a face on a particular position from the set every time, then when there were 75% happy faces in the display, there would be 75% chances that participants saw a happy face and 25% chances seeing an angry (by virtue of the trial by trial randomization of the positions occupied by angry and happy faces in the set). However, if this would be the case, then the distribution of face emotion ratings in the 75% happy faces condition should be bimodal in essence, namely, a large peak should fall at the upper side of the continuous 0-100 scale and another one, smaller in magnitude, at the lower side of the 0-100 scale. In contrast, if participants judged the average emotion based on a holistic processing the face set, and then the response distribution should be more uniform (or less predictable), with the position of the peak coinciding with the valence of the (perceived) mean emotion.

To formally assess a unimodal versus a bimodal distribution for the emotion ratings under each combination of ratio and condition, we relied on mixture modeling. More specifically, we tested the null hypothesis of a single normal distribution versus the hypothesis of a mixture of two normal distributions for Experiment 1 and Experiment 3 separately. If this null hypothesis was rejected, we next tested the null hypothesis of a mixture of two normal distributions versus the alternative of a mixture of three normal distributions. All those analyses were performed using the R-package mixtools (Benaglia, Chauveau, Hunter, & Young, 2009). The bootstrap is used in this package to construct the null distribution of the test statistics for the aforementioned null hypotheses. The package uses the EM-algorithm to estimate the means and standard deviations from the normal distributions, and the mixing proportions. Because independent observations were assumed in these analyses, we first removed subject-specific effects by

subtracting from every judgment the subject-specific mean under that condition. As a consequence, observations at the upper end of the 0-100 scale will fall at the “positive side” after this transformation, while observations at the lower end of the 0-100 scale will fall at the “negative side” after this transformation.

(insert Supplementary Figures 2 & 3 about here)

The mixture analyses revealed that the distribution of responses was always unimodal or bimodal, but never three-modal. The estimated means and standard deviations (denoted as *mu* and *sigma*) from the one- or two-component distributions and the mixing proportions (denoted as *lambda* in case of a two-component distribution) are shown in Supplementary Table 2.

Supplementary Figures 2 & 3 revealed the responses on the continuous scale were more dispersed and fell on the “wrong” side more frequently in the simultaneous condition compared with the sequential and the repeated conditions (especially in the more ambiguous ratio conditions), indicating that it was indeed more difficult to judge the average emotion in the simultaneous condition. In the simultaneous condition, there was each time a (second) sub-component of responses falling on the “wrong” side, however the fact that this pattern (i.e., the presence of two main components accounting for the response distribution) was similar across all five ratios but different for the sequential and repeated conditions suggested that the use of a systematic subsampling strategy was unlikely. However, based on this outcome (in the simultaneous condition), we could not formally exclude the possibility that participants selected one or several subsets of faces in the set to carry out the task. Hence, a parsimonious interpretation of these auxiliary results (response distribution) entails that participants in the simultaneous condition did likely use a subsampling strategy.

Supplementary Table 1A

Summary of Multilevel Models to Predict Average Emotion Judgments in Experiments 1

AIC\ Predictors	Type	Ratio	Type * Ratio	Mean Intensity	Variance Intensity
67039.7	$F(2, 7587) < 1$	$F(4, 7587) = 597.53$, $p < .001$	$F(8, 7587) = 2.55$, $p = .009$		
67006.3*	$F(2, 7586) < 1$	$F(4, 7586) = 20.52$, $p < .001$	$F(8, 7586) = 2.51$, $p = .010$	$F(1, 7586) = 37.50$, $p < .001$	
67008.4	$F(2, 7585) < 1$	$F(4, 7585) = 19.81$, $p < .001$	$F(8, 7585) = 2.48$, $p = .011$	$F(1, 7585) = 38.39$, $p < .001$	$F(1, 7585) < 1$

Supplementary Table 1B

Summary of Multilevel Models to Predict Average Emotion Judgments in Experiments 2

GCS\ Predictors	Type	Ratio	Type * Ratio	Mean Intensity	Variance Intensity
8082.21	$F(2, 8095) = 2.63$, $p = .072$	$F(4, 8095) = 348.77$, $p < .001$	$F(8, 8095) = 6.50$, $p < .001$		
8081.74*	$F(2, 8094) = 2.62$, $p = .073$	$F(4, 8094) = 41.08$, $p < .001$	$F(8, 8094) = 6.48$, $p = .001$	$F(1, 8094) = 1.99$, $p = .158$	
8081.84	$F(2, 8093) = 2.62$, $p = .073$	$F(4, 8093) = 41.09$, $p < .001$	$F(8, 8093) = 6.48$, $p = .001$	$F(1, 8093) = 1.98$, $p = .159$	$F(1, 8093) < 1$

Supplementary Table 1C

Summary of Multilevel Models to Predict Average Emotion Judgments in Experiments 3

AIC\ Predictors	Type	Ratio	Type * Ratio	Mean Intensity	Variance Intensity
70287.4	$F(2, 7887) = 1.46$, $p = .233$	$F(4, 7887) = 659.56$, $p < .001$	$F(8, 7887) = 3.46$, $p < .001$		
70285.8*	$F(2, 7886) = 1.48$, $p = .228$	$F(4, 7886) = 50.35$, $p < .001$	$F(8, 7886) = 3.47$, $p < .001$	$F(1, 7886) = 4.98$, $p = .026$	
70288.2	$F(2, 7885) = 1.47$, $p = .231$	$F(4, 7885) = 47.99$, $p < .001$	$F(8, 7885) = 3.45$, $p < .001$	$F(1, 7885) = 5.28$, $p = .022$	$F(1, 7885) <1$

Note. AIC (Akaike information criterion) and GCS (Generalized Chi-Square) are measures of the relative quality of different statistical models for the given continuous and binary data set respectively. The smaller the AIC and GCS, the better the model fits the data. The predictors include Type (simultaneous, sequential and repeated display type), Mean intensity (computed for each face display and for each participant based on their own emotional ratings), Variance intensity (the sample standard deviation calculated for each face display and for each participant), Ratio (25%, 37.5%, 50%, 62.5% and 75% happy faces in the set), and the interactions between Type and Mean Intensity, as well as the interaction between Type and Ratio. The star signal (*) highlights the smallest AIC or GCS (thus the best model).

Supplementary Table 2A

Summary of Parameters Obtained for Each Component in the Response Distribution in Experiment 1

Condition		*lambda*	*mu*	*sigma*
SI	0.25	0.566	-8.474	10.814
		0.434	11.042	21.287
	0.375	0.691	-9.543	14.560
		0.309	21.318	12.359
	0.5	0.414	-19.157	11.413
		0.586	13.508	15.050
	0.625	0.439	-18.916	15.262
		0.561	14.809	12.386
	0.75	0.190	-31.358	12.900
		0.810	7.346	13.701
SE	0.25	0.868	-3.914	12.610
		0.132	25.676	18.738
	0.375	0.879	-4.090	12.891
		0.121	29.801	10.546
	0.5	NA	0	18.444
	0.625	NA	0	18.137
	0.75	0.086	-30.190	17.252
		0.914	2.837	13.463
RE	0.25	0.684	-7.088	12.861
		0.316	15.365	18.303
	0.375	0.683	-8.256	13.297
		0.317	17.792	14.972

0.5	NA	0	19.092
0.625	NA	0	19.053
0.75	0.151	-29.752	10.099
	0.849	5.277	14.551

Supplementary Table 2B

Summary of Parameters Obtained for Each Component in the Response Distribution in Experiment 3

Condition		*lambda*	*mu*	*sigma*
SI	0.25	0.756	-8.351	11.172
		0.244	25.945	16.231
	0.375	0.547	-14.238	11.445
		0.453	17.177	15.392
	0.5	0.424	-21.003	11.223
		0.576	15.449	14.068
	0.625	0.295	-24.978	13.174
		0.705	10.428	16.239
	0.75	0.449	-15.388	18.895
		0.551	12.547	9.116
SE	0.25	0.893	-3.861	12.969
		0.107	32.289	11.421
	0.375	NA	0	19.988
	0.5	NA	0	19.565
	0.625	0.220	-25.491	13.833
		0.780	7.184	14.916
	0.75	0.171	-28.032	15.840
		0.829	5.778	13.888
RE	0.25	0.656	-8.728	11.205
		0.344	16.636	17.286
	0.375	0.617	-10.845	12.688
		0.383	17.482	16.608

0.5	NA	0	20.146
0.625	0.140	-27.931	15.068
	0.860	4.546	15.736
0.75	0.269	-21.735	17.841
	0.731	8.013	12.511

Note. The parameters were given by the function normalmixEM, separately for the five different ratios of happy faces and the three different display types. If the responses had one single normal distribution, the lambda was not available (NA). We used the sample mean and sample deviation as estimators for mu and sigma. SI = simultaneous; SE = sequential; RE = repeated.

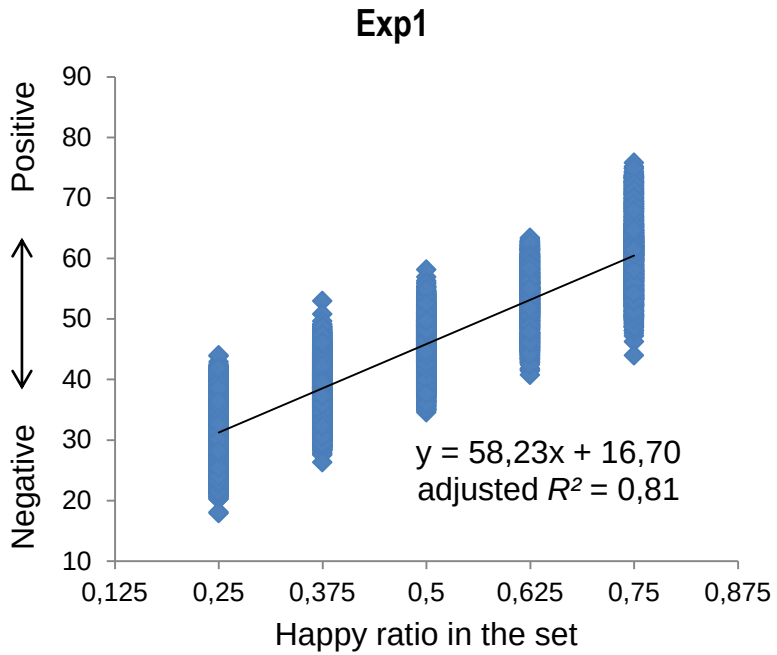
Supplementary Figure Legends

Figure 1. The computed mean intensity collapsed across display types shown separately for the five different ratios in Experiment 1 (A), Experiment 2 (B) and Experiment 3(C). The larger the value, the more positive the computed mean emotion intensity was, while conversely, the smaller this value, the more negative the computed mean intensity was. Significant linear effects were found for all three experiments, and the regression model was provided for each experiment separately.

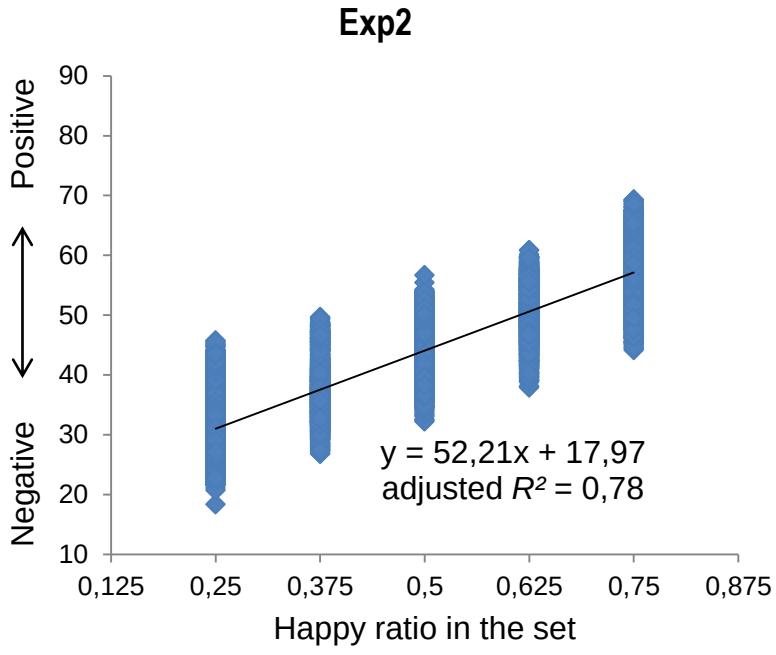
Figure 2. The distribution of average emotion judgements in the (A) simultaneous, (B) sequential, and (C) repeated conditions across five ratios in Experiment 1, given by the R-package mixtools.

Figure 3. The distribution of average emotion judgements in the (A) simultaneous, (B) sequential, and (C) repeated conditions across five ratios in Experiment 3, given by the R-package mixtools.

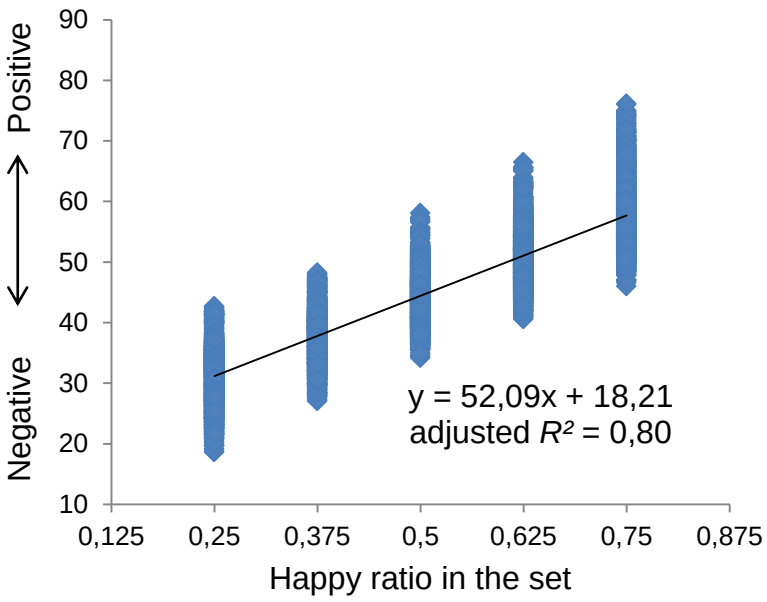
(A)

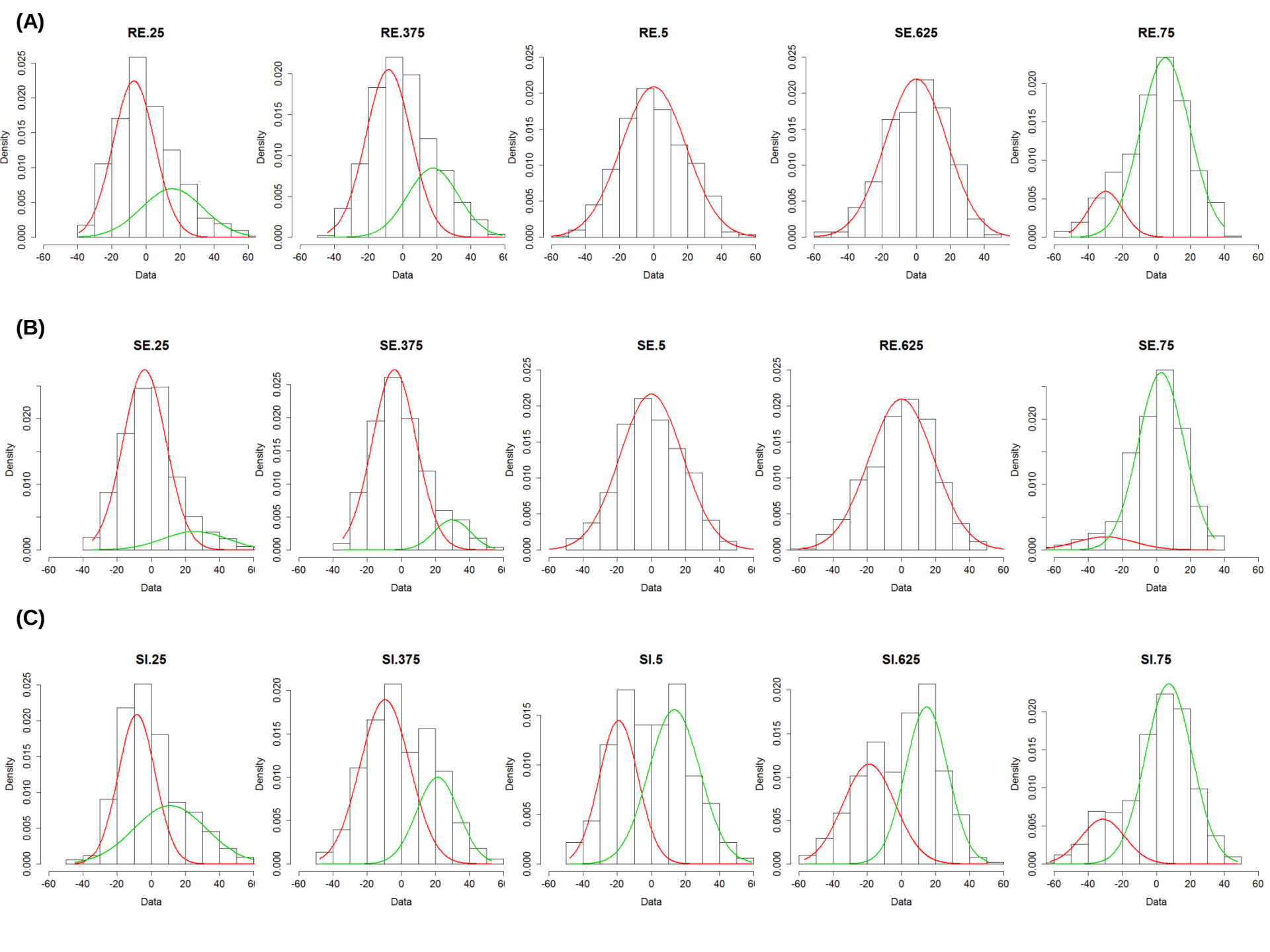

Exp1
Positive
Negative
90
80
70
60
50
40
30
20
10
0,125
0,25
0,375
0,5
0,625
0,75
0,875
y = 58,23x + 16,70
adjusted R² = 0,81
Happy ratio in the set

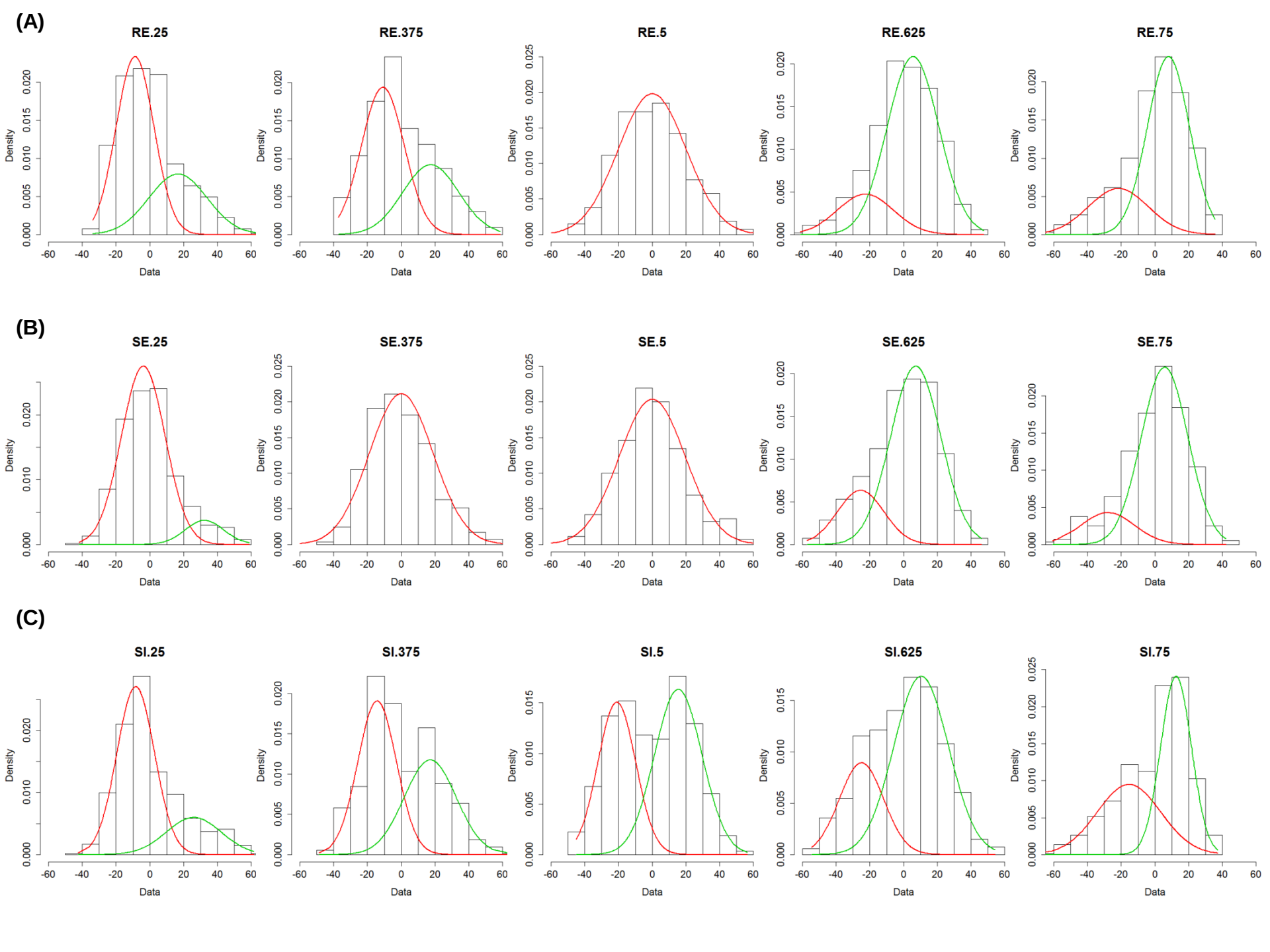
(B)


Exp2
Positive
Negative
90
80
70
60
50
40
30
20
10
0,125
0,25
0,375
0,5
0,625
0,75
0,875
y = 52,21x + 17,97
adjusted R² = 0,78
Happy ratio in the set

(C)


Exp3
Positive
Negative
90
80
70
60
50
40
30
20
10
0,125
0,25
0,375
0,5
0,625
0,75
0,875
y = 52,09x + 18,21
adjusted R² = 0,80
Happy ratio in the set


(A)
RE.25
RE.375
RE.5
SE.625
RE.75
(B)
SE.25
SE.375
SE.5
RE.625
SE.75
(C)
SI.25
SI.375
SI.5
SI.625
SI.75
Density
Data


(A)
RE.25
RE.375
RE.5
RE.625
RE.75
(B)
SE.25
SE.375
SE.5
SE.625
SE.75
(C)
SI.25
SI.375
SI.5
SI.625
SI.75
Density
Data